

# PP 价格月度预测

2025-2026 学年《数据挖掘与机器学习》课程个人 Big Project

姓名 (学号)

2026 年 1 月 10 日

(请将\StudentName 与\StudentID 替换为真实姓名与学号；最终提交 PDF 文件名按作业要求重命名)

## 目录

<b>1 问题背景与任务定义</b>	<b>3</b>
<b>2 数据说明与预处理</b>	<b>3</b>
2.1 数据来源与覆盖差异	3
2.2 统一降频到月度 (周/日 → 月)	3
2.3 目标变量与标签构造 (问题 1/2 共用)	4
2.4 涨跌强度分档 (问题 2)	4
2.5 缺失值处理与标准化	4
<b>3 特征工程与描述性统计 (时间序列 EDA)</b>	<b>4</b>
3.1 原始数据概览	4
3.2 月度特征覆盖与缺失	7
3.3 建模数据集 (pp_base) EDA	8
3.4 派生特征工程 (可选开关)	9
<b>4 问题 1: PP 价格月度预测与涨跌方向</b>	<b>9</b>
4.1 验证方案 (时间序列划分)	9
4.2 基线模型 (必须项)	9
4.3 机器学习模型与参数	9
4.4 结果与对比分析	10
<b>5 问题 2: 涨跌强度预测与概率输出</b>	<b>10</b>
5.1 建模路径	10
5.2 评价指标	11
5.3 结果摘要	11
<b>6 问题 3: 分阶段关键因子筛选与权重</b>	<b>11</b>
6.1 阶段划分 (行业周期)	11
6.2 关键因子组自动筛选	12
<b>7 总结: 不足与展望</b>	<b>12</b>

目录	2
附录目录	13
A 附录 A：可复现代码结构与运行命令	13
B 附录 B：主要输出文件说明	13
C 附录 C：完整模型对比表（CSV 路径）	13

## 1 问题背景与任务定义

数据集 PP 数据/ 包含企业搜集的 PP（聚丙烯）相关上下游指标（表 1-17）及汇总表（表 0），既有日度/周度也有月度数据，起止时间不完全一致。企业希望进行 **PP 价格的月度预测**。根据题目要求，本报告解决：

- **问题 1：**月度预测 PP 价格，并输出未来涨跌方向（涨/跌）。
- **问题 2：**预测涨跌强度（五分类示例：大涨/小涨/平/小跌/大跌），并输出涨跌幅度的数值预测与概率（predict\_proba）。
- **问题 3：**考虑行业周期分阶段，设计自动筛选关键因子组合的方法，并给出各阶段因子影响权重。

**符号约定（时间序列对齐）** 以月度为单位，记目标月价格为  $y_t$ ，特征为  $\mathbf{x}_t$ 。为避免信息泄露，本项目统一使用

$$\mathbf{x}_{t-1} \rightarrow y_t$$

即对所有特征做 shift(1)，用上一月信息预测下一月价格（预测步长  $t + 1$ ）。

## 2 数据说明与预处理

### 2.1 数据来源与覆盖差异

预测目标价格来自 PP 数据/1-华东市场 PP 粒市场价 \_ 法定工作日.xlsx。其余特征覆盖产量、进口、开工率、成本、库存、检修、GDP、期货等。由于频率与覆盖差异明显，需统一降频与缺失处理。表 1 给出各因子的覆盖概览（由 scripts/run\_pp\_eda.py 自动统计）。

表 1: 原始数据时间覆盖与频率概览（PP 数据/ 表 1-16）

因子	口径	起始日期	结束日期	频率判断	覆盖 (月)
BOPP 开工率	mma	2010-01-01	2021-06-01	月度	138
CTO 成本	mma	2015-01-04	2021-07-15	日度/不规则	79
GDP	single	2016-12-31	2020-12-31	年度/稀疏	5
MTO 成本	mma	2016-01-04	2021-07-15	日度/不规则	67
PDH 成本	mma	2019-02-13	2021-07-15	日度/不规则	30
PP 产量	mma	2014-01-31	2021-06-30	月度	90
PP 价格	mma	2014-12-01	2021-07-15	日度/不规则	80
PP 开工率	mma	2014-01-03	2021-07-15	周度	91
PP 石化库存	single	2011-01-16	2021-07-15	日度/不规则	127
PP 进口量	mma	2010-01-31	2021-05-31	月度	137
丙烯成本	mma	2011-01-04	2021-07-15	日度/不规则	127
乙烯成本	mma	2014-12-31	2021-07-15	日度/不规则	80
塑编开工率	mma	2011-01-07	2021-07-15	周度	127
排产比例	mma	2014-01-02	2021-07-15	日度/不规则	91
期货价格	mma	2019-10-28	2021-07-15	日度/不规则	22
检修损失	mma	2018-01-31	2021-06-30	月度	42

### 2.2 统一降频到月度（周/日 → 月）

对每一张原始表，统一生成月度特征口径：

- 若表中包含 最低/最高/平均：月内分别做 mean/min/max，并额外计算 last（该月最后一个观测日的平均）与 range（max-min）。

- 若表中为单值序列 (如库存、GDP): 同样计算 mean/min/max/last/range。

### 2.3 目标变量与标签构造 (问题 1/2 共用)

以 月均价口径为默认 (也支持月末价口径), 构造:

- 价格回归目标:  $y_t$  (月度 PP 价格)
- 上月价格:  $y_{t-1}$  (记为 `y_prev`)
- 方向标签 (严格口径): 若  $(y_t - y_{t-1}) > 0$  为涨 (1), 否则为跌/不涨 (0)
- 月度收益率:  $r_t = (y_t - y_{t-1})/y_{t-1}$  (记为 `y_return`)

### 2.4 涨跌强度分档 (问题 2)

默认阈值(可配置): 强波动阈值 5% (`strong_threshold=0.05`), 平稳阈值 0.5% (`flat_threshold=0.005`)。据此将  $r_t$  分为 5 类:

`big_down, small_down, flat, small_up, big_up`

并在分类模型中输出各档概率 (`predict_proba`)。

### 2.5 缺失值处理与标准化

建模 Pipeline 内对全部数值特征进行:

- 中位数填补 (median imputation) + 缺失指示特征
- 标准化 (StandardScaler)

保证不同量纲特征可比, 并将缺失信息纳入模型表达。

## 3 特征工程与描述性统计 (时间序列 EDA)

本节对应题目对“特征工程与描述性统计分析”的要求。我们同时对 原始数据、月度聚合后的宽表、以及 建模数据集 做统计与可视化 (输出均为 PDF)。

### 3.1 原始数据概览

我们对 PPData 的每张原始表都生成了“原始序列 + 月度平滑 + 分布直方图”(路径: `outputs/eda/raw/figu`) 用于检查频率差异、异常值与分布形态。为避免“图在上一页、解释在下一页”带来的空白, 本节采用“一行两图 + 紧接文字解读”的排版方式: 先展示同类因子 (两张图), 再分别解释每个因子的序列特征与业务含义。

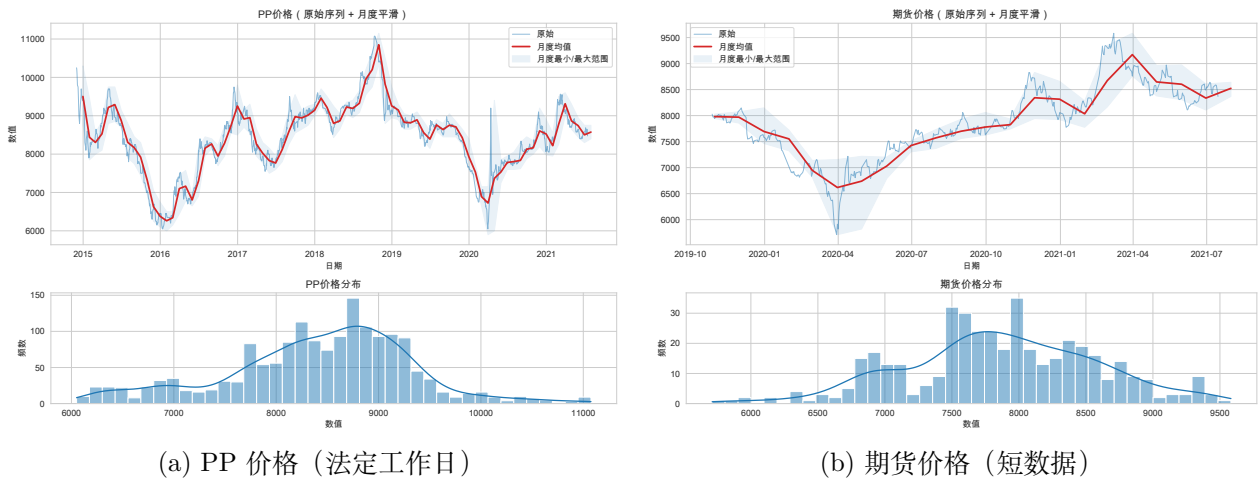


图 1: 原始数据: 价格相关因子

**PP 价格 (现货):** 覆盖 2014-12 至 2021-07 (约 80 个月), 日度/不规则数据; 均值约 8413、标准差约 907, 取值范围约为 [6050, 11075]。时序上呈现明显的周期波动与波动聚集: 2015–2016 年低位下行, 随后进入多轮上行与回落, 2018 年末出现阶段性高位, 2020 年初出现急跌后修复。图中浅色“月度最小/最大范围”在拐点与冲击月份显著变宽, 提示月内波动状态 (range、max-min) 本身携带有效信息; 分布上主峰集中在中高价区间且右尾更长, 说明极端高价月份虽少但会显著抬升均值并加大波动, 从而提高均值预测难度。

**期货价格:** 覆盖 2019-10 至 2021-07 (约 22 个月), 均值约 7866、标准差约 707, 范围约为 [5711, 9582], 且存在少量缺失 (约 1.4%)。时序上与现货同向且对拐点更敏感 (2020 年急跌、2021 年快速上行阶段更为明显); 在月度聚合层面, 期货与现货价格的斯皮尔曼相关在 lag0 约 0.97、lag1 仍约 0.89, 表明其对价格水平解释力很强。业务上期货更接近“预期/风险偏好”的快变量, 但覆盖期短会缩短有效样本, 因此我们保留“纳入/不纳入期货”两套方案并进行对比评估。

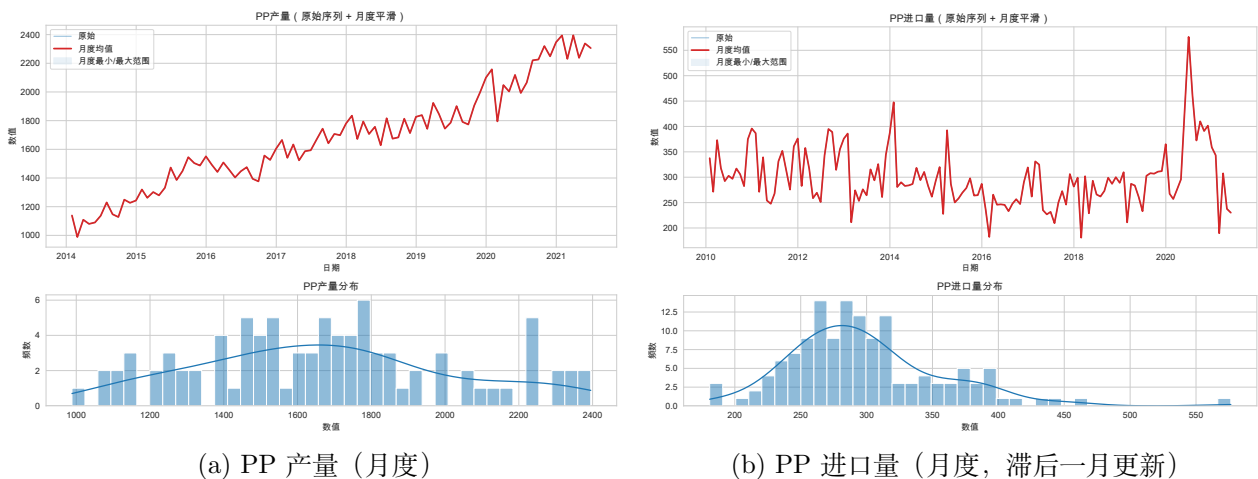


图 2: 原始数据: 供需相关因子

**PP 产量:** 覆盖 2014-01 至 2021-06 (90 个月), 月度统计; 均值约 1676、标准差约 348, 范围约为 [988, 2395]。序列整体趋势上行, 反映国内产能投放与装置稳定运行能力提升; 月度波动并非随机噪声, 往往对应检修、装置开停与需求旺淡季。由于产量具有趋势项, 其与价格的简单相关并不强 (例如月度聚合后 lag0 相关约 0.10、lag1 约 0.09), 说明“供给水平”对价格的影响往往需要与成本、库存、需求共同解释, 单独使用可能产生误判。

**PP 进口量:** 覆盖 2010-01 至 2021-05 (137 个月), 月度统计且源数据标注“滞后一月更新”; 均值

约 299、标准差约 59，范围约为 [181, 576]。分布明显右偏并存在尖峰（少数月份进口量显著高于常态），体现外盘资源、国际价差与到港节奏的集中性。月度聚合后其与价格在较长滞后处出现负相关（例如 lag6 约 -0.13），符合“高价吸引进口、到港存在运输滞后”的机制。建模上更适合使用相对量与变化率（如 供需 \_\_ 进口 \_div\_ 产量 \_\_mean、动量/滚动均值），并严格用上一月信息预测下一月以规避更新滞后带来的信息穿越。

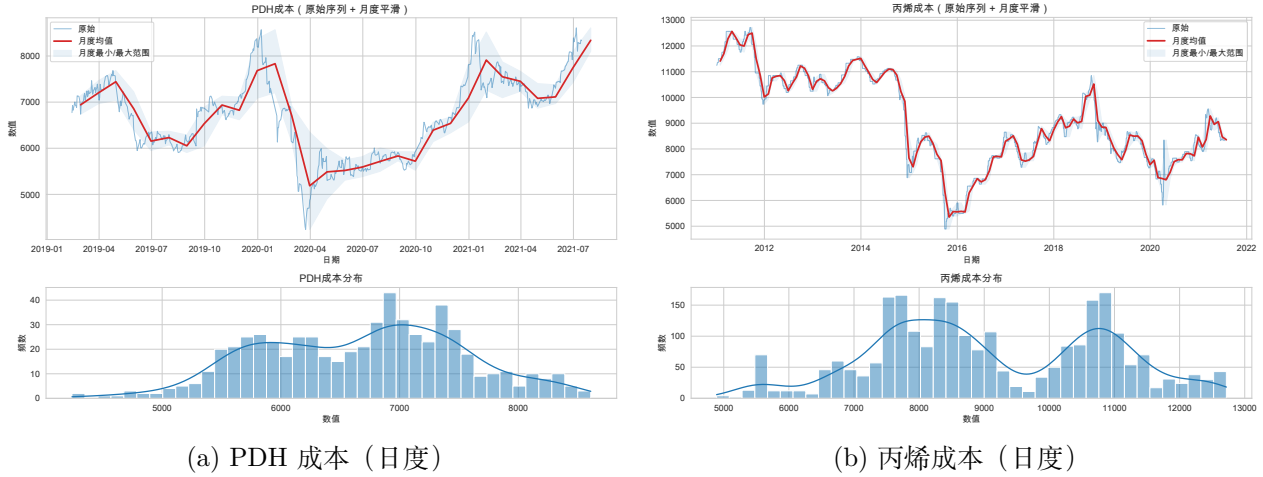


图 3: 原始数据: 成本相关因子

**PDH 成本:** 覆盖 2019-02 至 2021-07 (约 30 个月)，日度数据；均值约 6677、标准差约 848，范围约为 [4240, 8610]。2020 年前后存在明显的断崖式下跌与随后修复，体现能源/原料端冲击的快速传导。月度聚合后其与价格在短滞后处呈中等正相关 (lag0 约 0.46)，但由于覆盖期短，相关结构更容易受少数极端月份影响，需通过正则化、集成与稳健特征（价差/比值）控制不确定性。

**丙烯成本:** 覆盖 2011-01 至 2021-07 (约 127 个月)，日度数据；均值约 9082、标准差约 1760，范围约为 [4891, 12718]。时序中出现多次“制度切换”（不同宏观/能源状态下的成本区间），分布呈多峰与厚尾；与 PP 价格的相关性极强（月度聚合后 lag0 约 0.89、lag1 仍约 0.85），符合“成本驱动定价”的产业逻辑。建模上应重点刻画利润空间（现货-成本价差）与成本冲击的持续性（滚动均值/波动/动量），以提升对成本行情的识别。

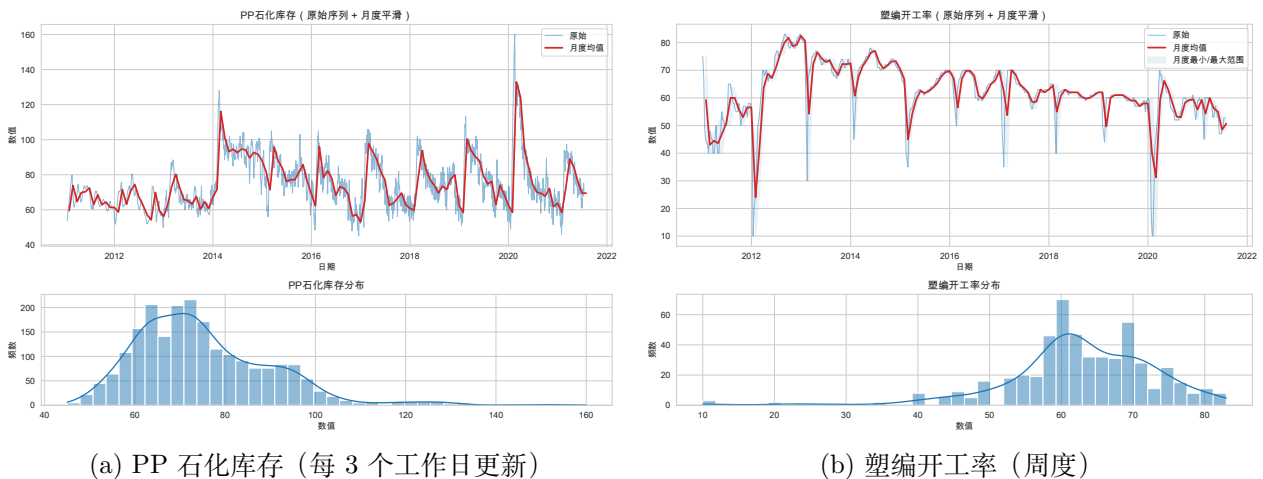


图 4: 原始数据: 库存与下游景气因子

**PP 石化库存:** 覆盖 2011-01 至 2021-07 (约 127 个月)，高频更新（每 3 个工作日）；均值约 75.1、标准差约 15.2，范围约为 [45, 160]。库存序列短期起伏明显且右尾较长，存在少数尖峰（累库冲击），说明库存更像“风险信号”：当库存快速上冲时，价格往往承压或进入震荡。月度聚合后其与价格在

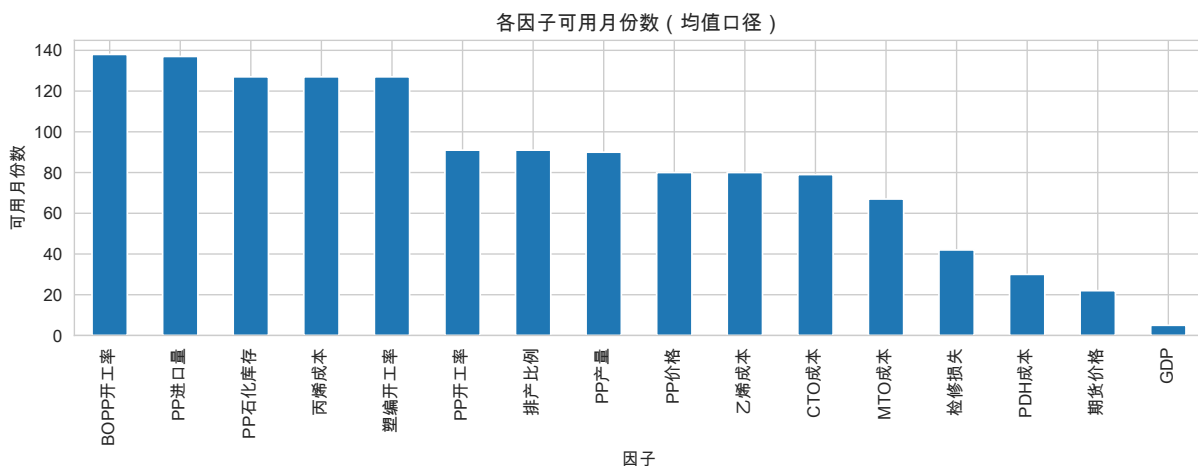
短滞后呈负相关 ( $\text{lag1}$  约 -0.17、 $\text{lag2}$  约 -0.22)，方向符合经验，但强度有限，提示其效应可能依赖成本与需求状态 (例如高成本下库存上升对价格的影响更大)。因此我们在特征工程中使用库存的  $\text{last/range}$  与比值特征来增强信号稳定性。

**塑编开工率**: 覆盖 2011-01 至 2021-07 (约 127 个月)，周度数据；均值约 62.6、标准差约 10.5，范围约为 [10, 83]。序列具有显著季节性 (年初/春节附近低谷) 并伴随少量异常低值点；长期中枢存在缓慢下移迹象，可能反映行业景气与产能利用率变化。其与价格的短滞后相关偏弱 ( $\text{lag0}$  约 -0.11)，但在更长滞后处相关有所增强 (例如  $\text{lag8}$  约 0.18)，提示下游景气对价格的影响可能存在传导滞后。建模上更适合通过滚动统计、动量与“下游开工指数”类聚合特征来表征需求持续性，而非仅用单月水平值。

### 3.2 月度特征覆盖与缺失



(a) 月度因子覆盖热力图 (1/0)



(b) 每个因子可用月份数量

图 5: 月度特征覆盖情况 (由 `scripts/run_pp_eda.py` 生成)

### 3.3 建模数据集 (pp\_base) EDA

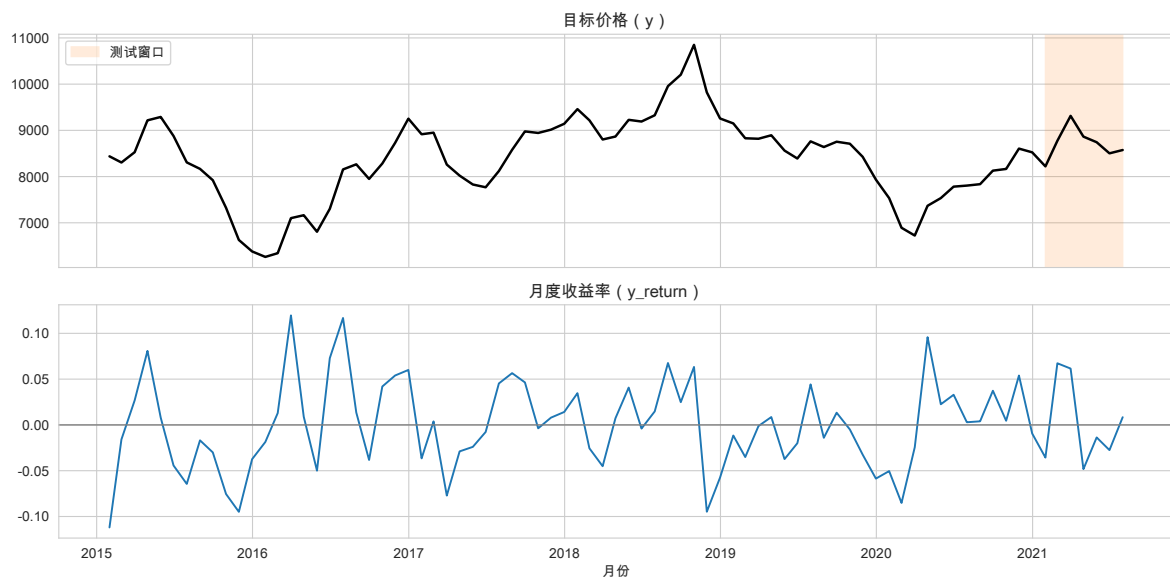
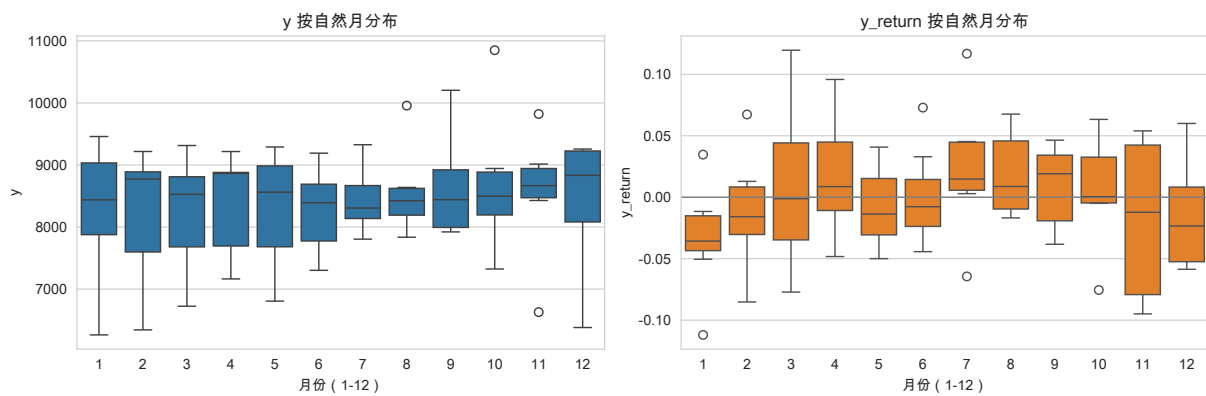
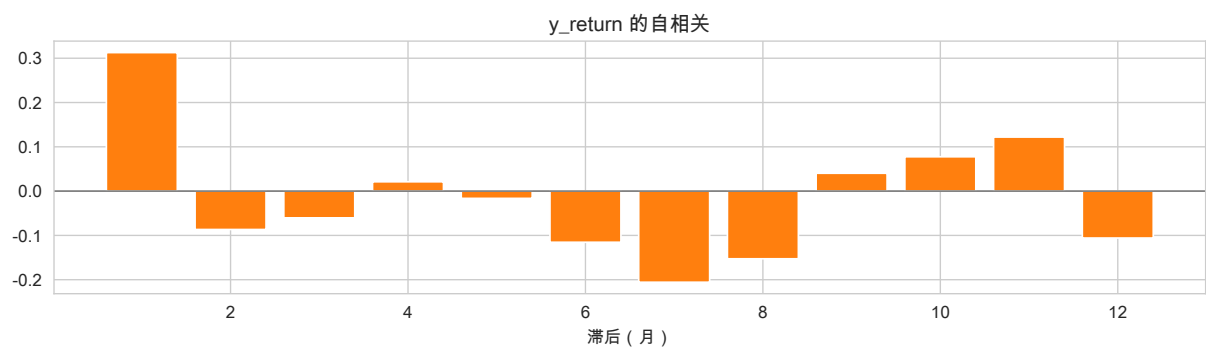


图 6: 目标序列:  $y$  与  $y\_return$  (含测试窗阴影)



(a) 按月份的季节性箱线图



(b) 收益率自相关 (ACF)

图 7: 季节性与自相关 (pp\_base)

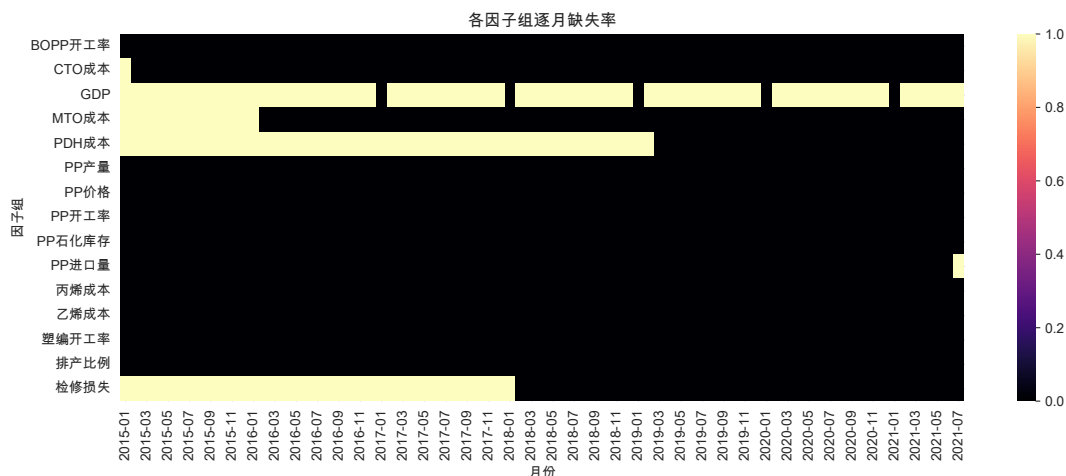


图 8: 按因子组统计的逐月缺失率热力图 (pp\_base)

### 3.4 派生特征工程 (可选开关)

在月度宽表基础上，额外构造滚动统计/动量/价差比值等派生特征（仅基于已滞后后的特征序列,避免使用未来信息）。该开关在数据集构建时通过 `--engineer-features` 启用,生成 `pp_base_engineered` 等数据集用于对比。

## 4 问题 1: PP 价格月度预测与涨跌方向

### 4.1 验证方案 (时间序列划分)

按题目提示，使用连续时间窗口作为测试集：默认测试窗为 **2021-01~2021-07**，训练集为其之前所有月份。评价指标包括：

- 回归：MAE、RMSE、MAPE
- 方向：Accuracy、Precision、Recall（方向定义严格按  $(y_t - y_{t-1}) > 0$  为涨）

### 4.2 基线模型 (必须项)

表 2: 问题 1 基线模型 (pp\_base, 测试窗 2021-01~2021-07)

Baseline	MAE	RMSE	MAPE(%)	Acc	Precision	Recall
naive ( $\hat{y}_t = y_{t-1}$ )	325.75	372.20	3.70	0.571	0.000	0.000
seasonal_12 ( $\hat{y}_t = y_{t-12}$ )	1336.20	1489.51	15.11	0.571	0.000	0.000

### 4.3 机器学习模型与参数

回归模型覆盖线性/非线性/集成方法，并提供若干集成学习策略（均值/中位数/截尾/训练集 CV 加权/Top-k）。主要模型与关键参数见表 3,完整参数会在运行后输出到 `outputs/metrics/<dataset>/pp_par`

表 3: 主要模型与参数设置 (节选)

模型	关键参数
Ridge	$\alpha = 1.0$
Lasso	$\alpha = 0.001$ , max_iter=20000
ElasticNet	$\alpha = 0.001$ , l1_ratio=0.5, max_iter=20000
BayesianRidge	默认参数
HuberRegressor	max_iter=2000, $\epsilon = 1.35$ , $\alpha = 1e-4$
KNNRegressor	n_neighbors=8, weights=distance
SVR(RBF)	$C = 10$ , $\gamma = \text{scale}$ , $\epsilon = 0.1$
RandomForestRegressor	n_estimators=500, max_depth=6, min_samples_leaf=2
ExtraTreesRegressor	n_estimators=1000, max_depth=8, min_samples_leaf=2
GBR	learning_rate=0.05, n_estimators=500, max_depth=3
AdaBoostRegressor	n_estimators=500, learning_rate=0.05
Bagging(Tree)	n_estimators=300, bootstrap=True, base_tree_depth=4
LogisticRegression(强度)	max_iter=2000
SVC(RBF, 强度)	$C = 5$ , $\gamma = \text{scale}$ , probability=True
RandomForestClassifier(强度)	n_estimators=500, max_depth=6, min_samples_leaf=2
ExtraTreesClassifier(强度)	n_estimators=1000, max_depth=8, min_samples_leaf=2
GBR_clf(强度)	learning_rate=0.05, n_estimators=500, max_depth=3
AdaBoostClassifier(强度)	n_estimators=500, learning_rate=0.05
Bagging(Tree, 强度)	n_estimators=300, bootstrap=True, base_tree_depth=4
SARIMAX(可选)	order=(1,1,1), 季节项 (1,0,1,12)(样本足够时启用), 外生变量取 Top-20 相关特征
Prophet(可选)	yearly_seasonality=True, additive; 外生回归量取 Top-10 相关特征

4.4 结果与对比分析

表 4 汇总了四种方案 (是否纳入期货、是否做额外特征工程) 在测试窗上的最优结果 (按 RMSE 选择)。

表 4: 问题 1: 价格点预测与涨跌方向 (最佳模型摘要, 按 RMSE 选取)

数据集	方案	最佳模型	MAE	RMSE	MAPI
pp_base	不含期货	ensemble_cv_weighted	158.05	193.36	
pp_base_engineered	不含期货 + 特征工程	ensemble_median	169.68	191.07	
pp_with_futures	含期货 (restrict)	huber	395.54	506.92	
pp_with_futures_engineered	含期货 + 特征工程 (restrict)	ensemble_topk_cv_weighted	277.81	359.46	

讨论

- **集成学习收益:** 在不含期货的长样本上, 训练集 CV 加权/中位数集成在 RMSE 上优于单一模型。
- **期货变量的敏感性:** 期货数据覆盖仅 22 个月, restrict 会削减可用样本; 因此“含期货”方案更适合做敏感性对比, 而非与长样本直接公平对照。
- **可解释性:** 线性模型 (Ridge/Huber) 便于解释系数与方向, 但在当前测试窗上, 树模型/集成表现更优。

5 问题 2: 涨跌强度预测与概率输出

5.1 建模路径

本项目同时提供两条实现路径:

- **路径 A (回归派生):** 先预测  $y_t$ , 再计算  $hatr_t = (haty_t - y_{t-1})/y_{t-1}$ , 并映射到强度档位 (得到数值预测)。
- **路径 B (强度多分类):** 直接以 `y_strength` 五分类建模, 输出每一档概率 `proba_*` (满足题目 “输出出现相关涨跌幅度的概率”)。

5.2 评价指标

对强度五分类, 报告: Accuracy、macro-Precision、macro-Recall、macro-F1 (时间窗同问题 1)。

5.3 结果摘要

表 5: 问题 2: 涨跌强度 (五分类) 与概率输出 (最佳模型摘要, 按 macro-F1 选取)

数据集	方案	最佳模型	F1_macro	Accuracy
pp_base	不含期货	ensemble_proba_topk_cv_weighted	0.467	0.42
pp_base_engineered	不含期货 + 特征工程	ensemble_proba_mean	0.600	0.57
pp_with_futures	含期货 (restrict)	logreg	0.242	0.57
pp_with_futures_engineered	含期货 + 特征工程 (restrict)	logreg	0.242	0.57

**概率输出文件** 对每个测试月, `pp_strength_test_predictions.csv` 与 `pp_strength_ensemble_test_predictions.csv` 给出 `strength_pred` 以及各档概率; 回归路径的数值涨跌幅预测见 `pp_test_predictions.csv` 的 `return_pred`。此外可选 `residual bootstrap` (`--bootstrap N`) 输出预测区间与  $p$ (`textup`)。

6 问题 3: 分阶段关键因子筛选与权重

6.1 阶段划分 (行业周期)

用月度收益率构造趋势与波动特征 (默认窗口 6 个月), 并按阈值/分位数划分高波动与趋势状态, 得到阶段标签 (`regime`), 再合并过短阶段形成最终阶段 (`stage_id`)。

表 6: 问题 3: 阶段划分结果摘要 (以 `pp_base` 为例)

阶段	市场状态 (regime)	起止月份	月数	累计涨跌 (%)	波动 (%/月)
1	flat_high_vol	2015-01 ~ 2015-08	8	-3.22	5.49
2	down_low_vol	2015-09 ~ 2016-02	6	-19.94	3.57
3	up_high_vol	2016-03 ~ 2017-03	13	16.32	5.99
4	up_low_vol	2017-04 ~ 2018-10	19	35.27	3.24
5	down_low_vol	2018-11 ~ 2020-03	17	-31.54	3.45
6	up_low_vol	2020-04 ~ 2021-07	16	16.37	3.87

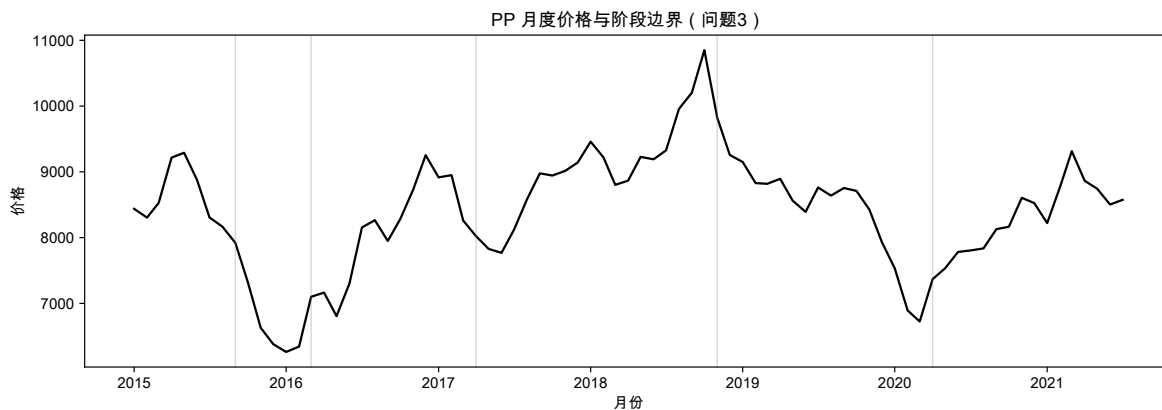


图 9: PP 月度价格与阶段边界 (pp\_base)

## 6.2 关键因子组自动筛选

在每个阶段内，以 **Ridge** 回归拟合

$\mathbf{y}_t$

并按“因子组”（同一前缀，如 丙烯成本 `__mean`）汇总绝对系数：

$$w(g) = \frac{\sum_f |\beta_f|}{\sum_f |\beta_f|}$$

取 Top-K 因子组作为该阶段关键因子组合。表 7 给出 pp\_base 的 Top-5 结果示例；强度任务同理使用多项 Logistic 回归的系数幅度作为权重（输出到 `q3_strength_logreg_group_weights.csv`）。

表 7: 问题 3: 各阶段关键因子组 (Top-5, Ridge 绝对系数归一化权重, pp\_base)

阶段	regime	Top-5 因子组 (权重)
1	flat_high_vol	丙烯成本 (0.177), PP 产量 (0.154), PP 石化库存 (0.142), 乙烯成本 (0.113), 排产比例 (0.109)
2	down_low_vol	CTO 成本 (0.161), 塑编开工率 (0.150), 丙烯成本 (0.148), 排产比例 (0.118), PP 开工率 (0.111)
3	up_high_vol	PP 石化库存 (0.230), 乙烯成本 (0.146), CTO 成本 (0.141), 排产比例 (0.102), PP 产量 (0.099)
4	up_low_vol	乙烯成本 (0.169), 塑编开工率 (0.144), PP 石化库存 (0.132), 丙烯成本 (0.111), 排产比例 (0.109)
5	down_low_vol	丙烯成本 (0.163), CTO 成本 (0.125), PP 产量 (0.095), PDH 成本 (0.089), 乙烯成本 (0.083)
6	up_low_vol	排产比例 (0.170), PDH 成本 (0.152), PP 石化库存 (0.144), PP 开工率 (0.127), CTO 成本 (0.125)

## 7 总结：不足与展望

- **数据覆盖不一致**：不同因子起止时间差异显著，尤其期货覆盖较短；未来可引入“缺失机制建模”或更精细的特征筛选策略。
- **样本量有限**：月度样本数不大，复杂模型易过拟合；可进一步采用滚动回测、多窗口稳定性检验。
- **概率输出改进**：当前强度概率来自分类器 soft-voting；未来可做概率校准 (Platt/Isotonic) 或分布预测 (分位数回归、block bootstrap)。
- **阶段划分改进**：可对比变点检测/HMM 等更贴近行业周期的分段方法，并结合业务解释阶段含义。

## 附录目录

- 附录 A: 可复现代码结构与运行命令
- 附录 B: 主要输出文件说明
- 附录 C: 完整模型对比表 (CSV 输出路径)

### A 附录 A: 可复现代码结构与运行命令

Listing 1: 生成数据集 / EDA / 建模 / 问题 3 (建议在 lchen 环境中运行)

```
conda run -n lchen python scripts/build_pp_dataset.py --target-metric mean --
    output outputs/datasets/pp_base.csv
conda run -n lchen python scripts/build_pp_dataset.py --include-futures --
    output outputs/datasets/pp_with_futures.csv

conda run -n lchen python scripts/build_pp_dataset.py --engineer-features --
    output outputs/datasets/pp_base_engineered.csv
conda run -n lchen python scripts/build_pp_dataset.py --include-futures --
    engineer-features --output outputs/datasets/pp_with_futures_engineered.csv

conda run -n lchen python scripts/run_pp_eda.py
conda run -n lchen python scripts/run_pp_models.py --dataset outputs/datasets/
    pp_base.csv

conda run -n lchen python scripts/run_q3_stage_factor_selection.py --dataset
    outputs/datasets/pp_base.csv --also-strength
```

### B 附录 B: 主要输出文件说明

- **数据集:** outputs/datasets/pp\_\*.csv
- **EDA:** outputs/eda/ (PDF 图 + CSV 统计表, 如 ACF、ADF、缺失率热力图等)
- **问题 1/2 指标:** outputs/metrics/<dataset>/pp\_model\_metrics.csv, pp\_strength\_model\_metrics.csv
- **问题 2 概率:** pp\_strength\_test\_predictions.csv, pp\_strength\_ensemble\_test\_predictions.csv
- **Bootstrap:** pp\_bootstrap\_predictions.csv (可选开关 --bootstrap)
- **问题 3:** outputs/q3/<dataset>/q3\_ridge\_group\_weights.csv, q3\_stages\_price.pdf 等

### C 附录 C: 完整模型对比表 (CSV 路径)

为控制正文页数, 完整模型对比明细以 CSV 形式保存于:

- outputs/metrics/pp\_base/pp\_model\_metrics.csv
- outputs/metrics/pp\_base/pp\_strength\_model\_metrics.csv
- 其余数据集同理 (目录名为数据集 stem)